

Distilling Information Reliability and Source Trustworthiness from Digital Traces

Behzad Tabibian¹, Isabel Valera², Mehrdad Farajtabar³, Le Song³,
Bernhard Schölkopf³, and Manuel Gomez-Rodriguez²

¹Max Planck Institute for Intelligent Systems, me@btabibian.com, bs@tue.mpg.de

²Max Planck Institute for Software Systems, ivalera@mpi-sws.org, manuelgr@mpi-sws.org

³Georgia Institute of Technology, mehrdad@gatech.edu, lsong@cc.gatech.edu

1 Introduction

Over the years, the Web has become a vast repository of information and knowledge about a rich variety of topics and real-world events – one much larger than anything we could hope to accumulate in conventional textbooks and traditional news media outlets. Unfortunately, due to its immediate nature, it also contains an ever-growing number of opinionated, inaccurate or false facts, urban legends and unverified stories of unknown or questionable origin, which are often refuted over time.^{1,2} To overcome this problem, online knowledge repositories, such as *Wikipedia*, *Stack Overflow* and *Quora*, put in place different evaluation mechanisms to increase the reliability of their content. These mechanisms can be typically classified as:

- I. **Refutation:** A user refutes, challenges or questions a statement contributed by another user or a piece of content originated from an external web source. For example, in *Wikipedia*, an editor can refute a questionable, false or incomplete statement in an article by removing it.
- II. **Verification:** A user verifies, accepts or supports a statement contributed by another user or a piece of content originated from an external web source. For example, in *Stack Overflow*, a user can accept or upvote the answers provided by other users.

However, these evaluation mechanisms only provide noisy measurements of the reliability of information and the trustworthiness of the information sources. Can we leverage these noisy measurements, often biased, to distill a robust, unbiased and interpretable measure of both notions?

In this paper, we argue that the *temporal traces* left by these noisy evaluations give cues on the reliability of the information and the trustworthiness of the sources. For example, while statements provided by an untrustworthy user will be often spotted by other users as unreliable and refuted quickly, statements provided by trustworthy users will be refuted less often. However, at a particular point in time, a statement about a complex, controversial or time evolving topic, story, or more generally, *knowledge item*, may be refuted by other users independently of the source. In this case, quick refutations will not reflect the trustworthiness of the source but the intrinsic unreliability of the knowledge item the statement refers to.

To explore this hypothesis, we propose a temporal point process modeling framework of refutation and verification in online knowledge repositories, which leverages the above mentioned temporal traces to obtain a meaningful measure of both information reliability and source trustworthiness. The key idea is to disentangle to what extent the temporal information in a statement evaluation (verification or refutation) is due to the intrinsic unreliability of the involved knowledge item, or to the trustworthiness of the source providing the

¹<http://www.snopes.com>

²<http://www.factcheck.org>

statement. To this aim, we model the times at which statements are added to a knowledge item as a counting process, whose intensity captures the temporal evolution of the reliability of the item—as a knowledge item becomes more reliable, it is less likely to be modified. Moreover, each added statement is supported by an information source and evaluated by the users in the knowledge repository at some point after its addition time. Here, we model the evaluation time of each statement as a survival process, which starts at the addition time of the statement and whose intensity captures both the trustworthiness of the associated source and the unreliability of the involved knowledge item.

For the proposed model, we develop an efficient method to find the optimal model parameters that jointly maximize the likelihood of an observed set of statement addition and evaluation times. This efficient algorithm allows us to apply our framework to ~ 19 million addition and refutation events in ~ 100 thousand *Wikipedia* articles and ~ 1 million addition and verification events in ~ 378 thousand questions in *Stack Overflow*. Our experiments show that our model accurately predicts whether a statement in a *Wikipedia* article (an answer in *Stack Overflow*) will be refuted (verified), it provides interpretable measures of source trustworthiness and information reliability, and yields interesting insights:³

- I. Most active sources are generally more trustworthy, however, trustworthy sources can be also found among less active ones.
- II. Changes on the reliability of a *Wikipedia* article over time, as inferred by our framework, match external noteworthy events.
- III. Questions and answers in *Stack Overflow* cluster into groups with similar levels of difficulty and popularity.

Related work. The research area most closely related to ours is on truth discovery and source trustworthiness. The former aims at resolving conflicts among noisy information published by different sources and the latter assesses the quality of a source by means of its ability to provide correct factual information. Most previous works have studied both problems together and measure the trustworthiness of a source using link-based measures [4, 17], information retrieval based measures [28], accuracy-based measures [8, 9, 29], content-based measures [2], and graphical model analysis [23, 30, 31, 32]. A recent line of work [20, 21, 22, 27] also considers scenarios in which the truth may change over time. However, previous work typically shares one or more of the following limitations, which we address in this work: (i) they only support knowledge triplets (subject, predicate, object) or structured knowledge; (ii) they assume there is a *truth*, however, a statement may be under discussion when a source writes about it; and, (iii) they do not distinguish between the unreliability of the knowledge item to which the statement refers and the trustworthiness of the source.

Temporal point processes have been previously used to model information cascades [15, 11, 5], social activity [14, 19, 12], badges [10], network evolution [18, 13], opinion dynamics [6], or product competition [26]. However, to the best of our knowledge, the present work is the first that leverages temporal point processes in the context of information reliability and source trustworthiness.

2 Background on Temporal Point Processes

A temporal point process is a stochastic process whose realization consists of a list of discrete events localized in time, $\{t_i\}$ with $t_i \in \mathbb{R}^+$ and $i \in \mathbb{Z}^+$. Many different types of data produced in social media and the Web can be represented as temporal point processes [6, 13, 26]. A temporal point process can be equivalently represented as a counting process, $N(t)$, which records the number of events up to time t , and can be characterized via its conditional intensity function — a stochastic model for the time of the next event given all the times of previous events. More formally, the conditional intensity function $\lambda^*(t)$ (intensity, for short) is given by

$$\lambda^*(t)dt := \mathbb{P}\{\text{event in } [t, t + dt) | \mathcal{H}(t)\} = \mathbb{E}[dN(t) | \mathcal{H}(t)],$$

³We will release an implementation of our inference method and datasets at <http://btabibian.com/projects/reliability>.

where $dN(t) \in \{0, 1\}$ denotes the increment of the process, $\mathcal{H}(t)$ denotes the history of event times $\{t_1, t_2, \dots, t_n\}$ up to but not including time t , and the sign $*$ indicates that the intensity may depend on the history. Then, given a time $t_i \geq t_{i-1}$, we can also characterize the conditional probability that no event happens during $[t_{i-1}, t_i]$ and the conditional density that an event occurs at time t_i as $S^*(t_i) = \exp(-\int_{t_{i-1}}^{t_i} \lambda^*(\tau) d\tau)$ and $f^*(t_i) = \lambda^*(t_i) S^*(t_i)$, respectively. Furthermore, we can express the log-likelihood of a list of events $\{t_1, t_2, \dots, t_n\}$ in an observation window $[0, T]$ as [1]

$$\mathcal{L} = \sum_{i=1}^n \log \lambda^*(t_i) - \int_0^T \lambda^*(\tau) d\tau. \quad (1)$$

This simple log-likelihood will later enable us to learn the parameters of our model from observed data. Finally, the functional form of the intensity $\lambda^*(t)$ is often designed to capture the phenomena of interests. Some useful functional forms we will use later are [1]:

- I. **Poisson process.** The intensity is assumed to be independent of the history $\mathcal{H}(t)$, but it can be a time-varying function, *i.e.*, $\lambda^*(t) = g(t) \geq 0$;
- II. **Hawkes Process.** The intensity models a mutual excitation between events, *i.e.*,

$$\lambda^*(t) = \mu + \alpha \sum_{t_i \in \mathcal{H}(t)} \kappa_\omega(t - t_i), \quad (2)$$

where $\kappa_\omega(t)$ is the triggering kernel, $\mu \geq 0$ is a baseline intensity independent of history. Here, the occurrence of each historical event increases the intensity by a certain amount determined by the kernel and the weight $\alpha \geq 0$, making the intensity history dependent and a stochastic process by itself; and,

- III. **Survival process.** There is only one event for an instantiation of the process, *i.e.*,

$$\lambda^*(t) = g(t)(1 - N(t)), \quad (3)$$

where $\lambda^*(t)$ becomes 0 if an event already happened before t and $g(t) \geq 0$.

3 Proposed Model

In this section, we formulate our modeling framework of verification and refutation in knowledge repositories, starting with the data representation it uses.

Data representation. The digital traces generated during the construction of a knowledge repository can be represented using the following three entities: the *statements*, which are associated to particular *knowledge items*, and the *information sources*, which support each of the statements. More specifically:

- An *information source* is an entity that supports a statement in a knowledge repository, *i.e.*, the web source an editor uses to support a paragraph in *Wikipedia*, the user who posts an answer on a Q&A site, or the software developer who contributes a piece of code in *Github*. We denote the set of information sources in a knowledge repository as \mathcal{S} .

- A *statement* is a piece of information contributed to a knowledge repository, which is characterized by its addition time t , its evaluation time τ , and the information source $s \in \mathcal{S}$ that supports it. Here, we represent each statement as the triplet

$$e = \begin{matrix} \text{source} & \text{evaluation time} \\ \downarrow & \downarrow \\ (s, & t, & \tau), \\ \uparrow & \\ \text{addition time} \end{matrix} \quad (4)$$

where an evaluation may correspond either to a verification or refutation.⁴ Moreover, if a statement is *never* refuted or verified, then we set $\tau = \infty$.

⁴For clarity, we assume that a knowledge repository either uses refutation or verification. However, our model can be readily extended to knowledge repositories using both.

— A *knowledge item* is a collection of statements. For example, a knowledge item corresponds to an article in *Wikipedia*; to a question and its answer(s) in a Q&A site; or to a software project on *Github*. Here, we gather the history of the d -th knowledge item, $\mathcal{H}_d(t)$, as the set of statements added to the knowledge item d up to but not including time t , *i.e.*,

$$\mathcal{H}_d(t) = \{e_i | t_i < t\}. \quad (5)$$

In most knowledge repositories, one can recover the source, addition time, and evaluation time of each statement added to a knowledge item. For example, in *Wikipedia*, there is an edit history for each *Wikipedia* article; on Q&A sites, all answers to a question are recorded; and, in *Github*, there is a version control mechanism to keep track of all changes.

Generative process for knowledge evolution. Our hypothesis is that the temporal information related to statement additions and evaluations reflects both the reliability of knowledge items and the trustworthiness of information sources. More specifically, our intuition is as follows:

- I. A reliable knowledge item should be stable in the sense that new statement addition will be rare, and it is less likely to be changed compared to unreliable items. Such notion of reliability should be reflected in the statement addition process—as a knowledge item becomes more reliable, the number of statement addition events within a unit of time should be smaller.
- II. A trustworthy information source should result in statements which are verified quickly and refuted rarely. Such notion of trustworthiness should be therefore reflected in its statement evaluation time—the more trustworthy an information source is, the shorter (longer) the time it will take to verify (refute) its statements.

In our temporal point process modeling framework, we build on the above intuition to account for both information reliability and source trustworthiness. In particular, for each knowledge item, we model the statement addition times $\{t_i\}$ as a counting process whose intensity directly relates to the reliability of the item—as a knowledge item becomes more reliable, it is less likely to be changed. Moreover, each addition time t_i is marked by its information source s_i and its evaluation time τ_i , which in turn depends on the source trustworthiness and also have an impact on the overall reliability of the knowledge item—the verification (refutation) of statements supported by trustworthy sources result in an increase (decrease) of the reliability of the knowledge item.

More in detail, for each knowledge item d , we represent the statement addition times $\{t_i\}$ as a counting process $N_d(t)$, which counts the number of statements that have been added up to but not including time t . Thus, we characterize the statement addition process using its corresponding intensity $\lambda_d^*(t)$ as

$$\mathbb{E}[dN_d(t) | \mathcal{H}_d(t)] = \lambda_d^*(t)dt, \quad (6)$$

which captures the evolution of the reliability of the knowledge item over time. Here, the smaller the intensity $\lambda^*(t)$, the more reliable the knowledge item at time t . Moreover, since a knowledge item consists of a collection of statements, the overall reliability of the knowledge item will also depend on the individual reliability of its added statements through their evaluations—the verification (refutation) of statements may result in an increase (decrease) of the reliability of the knowledge item, leading to an inhibition (increase) in the intensity of the statement additions to the knowledge item.

Additionally, every time a statement i is added to the knowledge item d , the corresponding information source $s_i \in \mathcal{S}$ is sampled from a distribution $p(s|d)$ and the evaluation time τ_i is sampled from a survival process, which we represent as a binary counting process $N_i(t) \in \{0, 1\}$, in which $t = 0$ corresponds to the time in which the statement is added and becomes one when $t = \tau_i - t_i$. Here, we characterize this survival process using its corresponding intensity $\mu_i^*(t)$ as

$$\mathbb{E}[N_i(t) | \mathcal{H}_d(t)] = \mu_i^*(t)dt, \quad (7)$$

which captures the temporal evolution of the reliability of the i -th statement added to the knowledge item. Here, the smaller the intensity $\mu_i^*(t)$, the shorter (longer) time it will take to verify (refute) it. This intensity

will depend, on the one hand, on the current intrinsic reliability of the corresponding knowledge item and, on the other hand, on the trustworthiness of the source supporting the statement.

Next, we formally define the functional form of the intensities $\lambda_d^*(t)$ and $\mu_i^*(t)$, and the source distribution $p(s|d)$.

Knowledge item reliability. For each knowledge item d , we consider the following form for its reliability function, or equivalently, its statement addition intensity:

$$\lambda_d(t) = \underbrace{\sum_j \phi_{d,j} k(t - t_j)}_{\text{item intrinsic reliability}} + \underbrace{\sum_{e_i \in \mathcal{H}_d(t)} \mathbf{w}_d^\top \gamma_{s_i} g(t - \tau_i)}_{\text{effect of past evaluations}}. \quad (8)$$

In the above expression, the first term is a mixture of kernels $k(t)$ accounting for the temporal evolution of the intrinsic reliability of a knowledge item over time, and the second term accounts for the effect that previous statement evaluations have on the overall reliability of the knowledge item. Here, \mathbf{w}_d and γ_{s_i} are L -length vectors whose elements indicate, respectively, the weight (presence) of each topic in the knowledge item and the per-topic influence of past evaluations of statements backed by source s_i . Finally, the function $g(t)$ is a nonnegative triggering kernel, which models the decay of the influence of past evaluations over time. If the evaluation is a refutation then we assume $\gamma_{s_i} \geq 0$, since a refuted statement typically decreases the reliability of the knowledge item and thus triggers the arrival of new statements to replace it. If the evaluation is a verification, we assume $\gamma_{s_i} \leq 0$, since a verified statement typically increases the reliability of the knowledge item and thus inhibits the arrival of new statements to the knowledge item. As a consequence, the above design results in an ‘‘evaluation aware’’ process, which captures the effect that previous statement evaluations exert on the reliability of a knowledge item.

Statement reliability. As discussed above, every statement addition event e_i is *marked* with an evaluation time τ_i , which we model using a survival process. The process is ‘‘statement driven’’ since it starts at the time when the statement addition event occurs and, within the process, $t = 0$ corresponds to the addition time of the statement. For each statement i , we adopt the following form for the statement reliability or, equivalently, for the intensity associated with its survival process:

$$\mu_i(t) = (1 - N_i(t)) \left[\underbrace{\sum_j \beta_{d,j} k(t + t_i - t_j)}_{\text{item intrinsic reliability}} + \underbrace{\mathbf{w}_d^\top \alpha_{s_i}}_{\substack{\text{source} \\ \text{trustworthiness}}} \right]. \quad (9)$$

In the above expression, the first term is a mixture of kernels $k(t)$ accounting for the temporal evolution of the intrinsic reliability of the corresponding knowledge item d and the second term captures the trustworthiness of the source that supports the statement. Here, \mathbf{w}_d and α_{s_i} are L -length nonnegative vectors whose elements indicate, respectively, the weight (presence) of each topic in the knowledge item d and the trustworthiness of source s_i in each topic. Since the elements \mathbf{w}_d sum up to one, the product $\mathbf{w}_d \alpha_{s_i}$ can be seen as the average trustworthiness of the source s_i in the knowledge item d . With this modeling choice, the higher the parameter α_{s_i} , the quicker the evaluation of the statement. Then, if the evaluation is a refutation, a high value of α_{s_i} implies low trustworthiness of the source s_i . In contrast, if it is a verification, a high value of α_{s_i} implies high trustworthiness.

Finally, note that the reliability of a statement, as defined in Eq. 9, reflects how quickly (slowly) it will be refuted or verified, and the reliability of a knowledge item, as defined in Eq. 8, reflects how quickly (slowly) new statements are added to the knowledge item.

Selection of source. The source popularity $p(s|d)$ typically depends on the topics contained in the knowledge item d . Therefore, we consider the following form for the source distribution:

$$p(s|d) = \sum_{\ell=1}^L w_{d,\ell} p(s|\ell), \quad (10)$$

where $w_{d,\ell}$ denotes the weight of topic ℓ in knowledge item d and $p(s|\ell) \propto \text{Multinomial}(\boldsymbol{\pi}_\ell)$ is the distribution of the sources for topic ℓ , *i.e.*, the vector $\boldsymbol{\pi}_\ell$ contains the probability of each source to be assigned to a topic ℓ .

4 Efficient Parameter Estimation

In this section, we show how to efficiently learn the parameters of our model, as defined by Eqs. 8 and 9, from a set of statement addition and evaluation events. Here, we assume that the topic weight vectors \mathbf{w}_d are given⁵. More specifically, given a set of sources \mathcal{S} and a set of knowledge items \mathcal{D} with histories $\{\mathcal{H}_1(T), \dots, \mathcal{H}_{|\mathcal{D}|}(T)\}$, spanning a time period $[0, T]$, we find the model parameters $\{\boldsymbol{\pi}_\ell\}_{\ell=1}^L$, $\{\boldsymbol{\beta}_d\}_{d=1}^{|\mathcal{D}|}$, $\{\boldsymbol{\phi}_d\}_{d=1}^{|\mathcal{D}|}$, $\{\boldsymbol{\alpha}_s\}_{s=1}^{|\mathcal{S}|}$ and $\{\boldsymbol{\gamma}_s\}_{s=1}^{|\mathcal{S}|}$, by solving the following maximum likelihood estimation (MLE) problem

$$\begin{aligned} & \text{maximize } \mathcal{L}(\{\boldsymbol{\pi}_\ell\}, \{\boldsymbol{\beta}_d\}, \{\boldsymbol{\phi}_d\}, \{\boldsymbol{\alpha}_s\}, \{\boldsymbol{\gamma}_s\}) \\ & \text{subject to } \boldsymbol{\pi}_\ell \geq 0, \boldsymbol{\beta}_d \geq 0, \boldsymbol{\phi}_d \geq 0, \boldsymbol{\alpha}_s \geq 0, \mathbf{1}^T \boldsymbol{\pi}_\ell = 1 \end{aligned}$$

where the log-likelihood is given by

$$\begin{aligned} \mathcal{L} = & \sum_{d=1}^{|\mathcal{D}|} \sum_{i: e_i \in \mathcal{H}_d(T)} \underbrace{\log p(t_i | \mathcal{H}_d(t_i), \boldsymbol{\phi}_d, \{\boldsymbol{\gamma}_s\}, \mathbf{w}_d)}_{\text{statements additions}} + \sum_{d=1}^{|\mathcal{D}|} \sum_{i: e_i \in \mathcal{H}_d(T)} \underbrace{\log p(\Delta_i | t_i, \boldsymbol{\beta}_d, \{\boldsymbol{\alpha}_s\}, \mathbf{w}_d)}_{\text{statements evaluations}} \\ & + \sum_{d=1}^{|\mathcal{D}|} \sum_{i: e_i \in \mathcal{H}_d(T)} \underbrace{\log p(s_i | \{\boldsymbol{\pi}_\ell\}, \mathbf{w}_d)}_{\text{sources popularity}}. \end{aligned} \quad (11)$$

In the above likelihood, the first term accounts for the times at which statements are added to the knowledge item, the second term accounts for the times at which statements are evaluated, and the third term accounts for the probability that source s_i is assigned to the statement addition event e_i . Since the first two terms correspond to likelihoods of temporal point processes, they can be computed using Eq. 1. The third term is simply given by $p(s_i | \{\boldsymbol{\pi}_\ell\}_{\ell=1}^L, \mathbf{w}_d) = \sum_{\ell=1}^L w_{d,\ell} \boldsymbol{\pi}_\ell(s_i)$, where $\boldsymbol{\pi}_\ell(s_i)$ denotes the s_i -th element of $\boldsymbol{\pi}_\ell$.

Remarkably, the above terms can be expressed as linear combinations of logarithms and linear functions or compositions of linear functions with logarithms and thus easily follow that the above optimization problem is jointly convex in all the parameters. Moreover, the problem can be decomposed into three independent problems, which can be solved in parallel obtaining local solutions that are in turn globally optimal. For knowledge repositories using refutation, *i.e.*, $\gamma_s \geq 0$, we solve both the first and second problem by adapting the algorithm by Zhou et al. [33]. For knowledge repositories using verification, *i.e.*, $\gamma_s \leq 0$, we solve the first problem using cvxpy [7] and the second problem by adapting the algorithm by Zhou et al. [33]. In both cases, the third problem can be computed analytically as

$$\boldsymbol{\pi}_\ell(s) = \frac{\sum_{d=1}^{|\mathcal{D}|} \mathbf{w}_d(\ell) \hat{\boldsymbol{\pi}}_d(s)}{\sum_{d=1}^{|\mathcal{D}|} \sum_{s'=1}^{|\mathcal{S}|} \mathbf{w}_d(\ell) \hat{\boldsymbol{\pi}}_d(s')}, \quad (12)$$

where $\mathbf{w}_d(\ell)$ denotes the ℓ -th element of \mathbf{w}_d , and $\hat{\boldsymbol{\pi}}_d(s)$ is the probability that source s is assigned to a statement in knowledge item d . In particular, $\hat{\boldsymbol{\pi}}_d(s)$ can be computed as

$$\hat{\boldsymbol{\pi}}_d(s) = \frac{n_{d,s}}{\sum_{s'=1}^{|\mathcal{S}|} n_{d,s'}}, \quad (13)$$

where $n_{d,s}$ is the number of statement addition events in the history of the knowledge item that are backed by source s , *i.e.*, $|\{e_i \in \mathcal{H}_d(T) | s_i = s\}|$. In practice, we found that adding a ℓ -1 penalty term on the parameters $\{\boldsymbol{\beta}_d\}$, *i.e.*, $\eta \sum_d \|\boldsymbol{\beta}_d\|_1$, which we set by cross-validation, avoids overfitting and improves the predictive performance of our model.

5 Experiments on Synthetic Data

Our goal in this section is to investigate if our parameter estimation method can accurately recover the true model parameters from statement addition and evaluation events. We examine this question using a synthetically generated dataset from our probabilistic model.

⁵There are many topic modeling tools to learn the topic weight vectors \mathbf{w}_d .

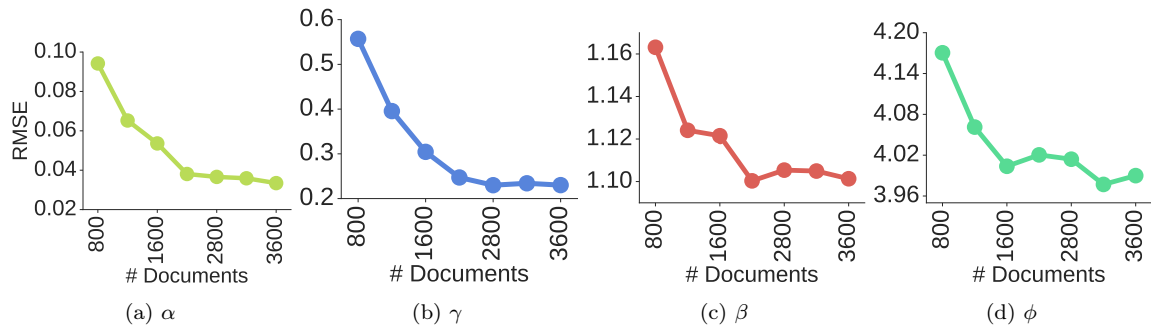


Figure 1: Performance of our model parameter estimation method on synthetic data in terms of root mean squared error (RMSE). The estimation becomes more accurate as we feed more knowledge items into our estimation procedure. However, since each new knowledge item increases the number of β and ϕ parameters, once the source parameter estimation becomes accurate enough, the estimation error for β and ϕ flattens.

Experimental setup. We set the number of sources to $|\mathcal{S}| = 400$, the total number of knowledge items to $|\mathcal{D}| = 3,600$, and assume the evaluation mechanism is refutation. We assume there is only one topic and then, for each source, we sample its trustworthiness α_s from the Beta distribution $Beta(2.0, 5.0)$ and its parameter γ_s from the uniform distribution $U(0, b)$, where $b = 0.03 \times \max(\{\alpha_s\}_{s \in \mathcal{S}})$. For the temporal evolution of the intrinsic reliability in the addition and evaluation processes, we consider a mixture of three radial basis (RBF) kernels located at times $t_j = 0, 6, 12$, with standard deviations of 2 and 0.5, respectively. Then, for each knowledge item, we first pick one of the kernel locations j uniformly at random, which determines the only *active* kernel for both the addition and the evaluation processes in the knowledge item, and sample their associated parameters, $\phi_{d,j}$ and $\beta_{d,j}$, from the log-normal distribution $\ln \mathcal{N}(3.5, 0.1)$ and the uniform distribution $U(0, 0.2\phi_d)$, respectively. Moreover, we assume that only up to five (different) sources are active in each knowledge item, which we pick at random, and then draw a source probability vector for these five active sources in the knowledge item from a Dirichlet distribution with parameter 0.5. The choice of prior distributions for the model parameters ensures enough variability across knowledge items and sources, so that the model parameters can be recovered. Finally, we generate addition and refutation samples from the resulting addition and evaluation processes during the time interval $(0, 15]$.

Results. We evaluate the accuracy of our model estimation procedure by means of the root mean square error (RMSE) between the true (x) and the estimated (\hat{x}) parameters, *i.e.*, $RMSE(x) = \sqrt{\mathbb{E}[(x - \hat{x})^2]}$. Figure 1 shows the parameter estimation error with respect to the number of knowledge items used to estimate the model parameters. Since the source parameters α and γ are shared across knowledge items, the estimation becomes more accurate as we feed more knowledge items into our estimation procedure. However, every time we observe a new knowledge item, the number of parameters increases with an additional β_d and ϕ_d . Therefore, the knowledge item parameter estimation only becomes more accurate as a consequence of a better estimation of the source parameters. As soon as the source parameter estimation becomes *good enough*, the estimation does not improve further and the estimation error flattens.

6 Experiments on Real Data

In this section, we apply our model estimation method to large-scale data gathered from two knowledge repositories: *Wikipedia*, which uses refutation as evaluation mechanism (*i.e.*, deleted statements), and *Stack Overflow*, which uses verification (*i.e.*, accepted answers). First, we show that our model can accurately predict whether a particular statement in a *Wikipedia* article will be refuted after a certain period of time, as well as which of the answers to a question in *Stack Overflow* will be accepted. Then, we show that it provides meaningful measures of web source trustworthiness in *Wikipedia* and user trustworthiness in *Stack*

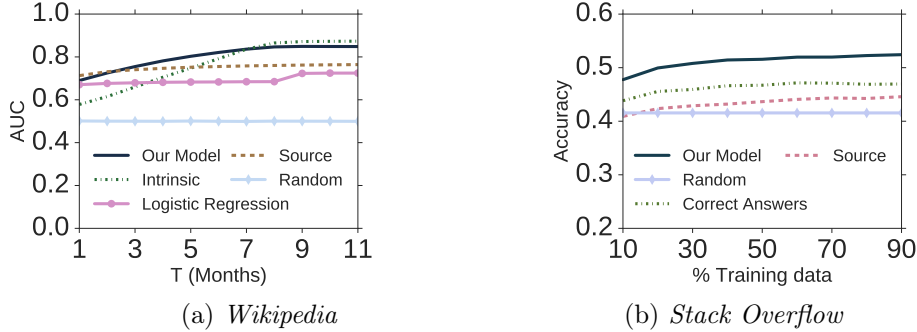


Figure 2: Prediction performance. Panel (a) shows the AUC achieved by our model and four baselines (Intrinsic, Source and Logistic Regression) for predicting whether a statement will be removed (refuted) from a *Wikipedia* article within a time period of T after it is posted; for different values of T . Panel (b) shows the success probability achieved by our model and two baseline (Source and Correct Answer) at predicting which answer to a question, among several answers, will be eventually verified in *Stack Overflow*.

Overflow. Finally, we demonstrate that our model can be used to: (i) pinpoint the changes on the intrinsic reliability of a *Wikipedia* article over time and these changes match external noteworthy controversial events; and, (ii) find question and answers in *Stack Overflow* with similar levels of intrinsic reliability, which in this case correspond to popularity and difficulty.

Data description and methodology. To build our *Wikipedia* dataset, we gather complete edit history, up to July 8, 2014, for 1 million *Wikipedia* English articles and track all the references (or links) to sources within each of the edits. Then, for each article d , we record for each added statement, its associated source s_i , its addition time t_i , and its refutation (deletion) time τ_i , if any. Such recorded data allows us to reconstruct the history of each article (or knowledge item), as given by Eq. 5. Moreover, since we can only expect our model estimation method to provide reliable and accurate results for articles and web sources with enough number of events, we only consider articles with at least 20 link additions and web sources that are used in at least 10 references. After these preprocessing steps, our dataset consists of ~ 50 thousand web sources that appeared in ~ 100 thousand articles, by means of ~ 10.4 million addition events and ~ 9 million refutation (deletion) events. The significant drop in the number of articles can be attributed to the large number of incomplete articles on Wikipedia, which lack reasonable number of citations. Finally, we run the (Python library) Gensim [24] on the latest revision of all documents in the dataset, with 10 topics and default parameters, to obtain the topic weight vectors \mathbf{w}_d , and apply our model estimation method, described in Section 4. Both in Eqs. 8 and 9, we used 19 RBF kernels, spaced every 9 months with standard deviation of 3 months. In Eq. 8, we used exponential triggering kernels with $\omega = 0.5 \text{ hours}^{-1}$.

To build our *Stack Overflow* dataset, we gathered history of answers from January 1, 2011 up to June 30, 2011, for ~ 500 thousand questions. Then, for each answer, we record the question d it belongs to, the user s_i who posted the answer, its addition time t_i , and its verification (acceptance) time τ_i , if any. Similarly as in the *Wikipedia* dataset, such recorded data allows us to reconstruct the history of each question (or knowledge item), as given by Eq. 5. Again, since our model estimation method can only provide reliable and accurate results for questions and users with enough number of events, we only consider questions with an accepted answer (if any exist) within 4 days of publication time and users who posted at least 4 accepted answers. After these preprocessing steps, our data consists of ~ 378 thousand questions which accumulate ~ 724 thousand addition events (answers) and ~ 224 thousand verification events (accepted answers). In this case, we assume a single topic and therefore the weight vector \mathbf{w}_d becomes a scalar value of 1. Finally, we apply our model estimation method, described in Section 4. In this case, in Eqs. 8 and 9, we used single constant kernels β_d and ϕ_d , respectively, since the intrinsic reliability of questions in *Stack Overflow* does not typically change over time. In Eq. 8, we used step functions as triggering kernels, since the inhibiting effect of an accepted answer does not decay over time.

| Music | | | Politics | |
|-------|------------------|------------------------|----------------|-----------------------|
| Rank | domain | Pr. rm. in 6 months | domain | Pr rm. in 6 months |
| 1 | guardian.co.uk | 0.15 | nytimes.com | 0.18 |
| 2 | rollingstone.com | 0.17 | guardian.co.uk | 0.19 |
| 3 | nytimes.com | 0.17 | google.com | 0.20 |
| 6 | billboard.com | 0.26 | usatoday.com | 0.24 |
| 13 | mtv.com | 0.32 | whitehouse.gov | 0.29 |
| Last | twitter.com | 0.56 | cia.com | 0.45 |

Table 1: Top 20 most popular web sources from *Wikipedia* in each topic ranked by the probability that a link from them is removed within 6 months (Most reliable on top).

In both datasets, our parameter estimation method runs in ~ 4 hours using a single machine with 10 cores and 64 GB RAM.

Can we predict if a statement will be removed from Wikipedia? Our model can answer this question by solving a binary classification problem: predict whether a statement will be removed (refuted) within a time period of T after it is posted.

— *Experimental setup:* We first split all addition events into a training set (90% of the data) and a test set (the remaining 10%) at random, then fit the parameters of the information survival processes given by Eq. 9 using only the evaluation times of the addition events from the training set, and finally predict whether particular statements in the test set will be removed within a time period of T after it is posted. We compare the performance of our model with three baselines: “Intrinsic”, “Source” and “Logistic Regression.” “Intrinsic” attributes all changes in an article to the intrinsic (un)reliability of that document. We can capture this assumption in our model by assuming that the parameter α_s in Eq. 9 is set to zero. Inspired by the model proposed by Adler and De Alfaro [2], we implement the baseline “Source”, which only accounts for the trustworthiness of the source that supports a statement, *i.e.*, it assumes that the intrinsic reliability of the article, parametrized by β_d in Eq. 9, is set to zero. Finally, “Logistic Regression” is a logistic regression model that uses the source identity (in one-hot representation), the document topic vector and the addition time of links as features. Here, we train a different logistic regression model per time window.

— *Results:* Since the dataset is highly unbalanced (only 25% of statements in the test set survive longer than 6 months), we evaluate the classification accuracy in terms of the area under the ROC curve (AUC), a standard metric for quantifying classification performance on unbalanced data. Figure 2(a) shows the AUC achieved by our model and the baselines for different values of T . Our model always achieves AUC values over 0.69, it improves its performance as T increases, and outperforms all baselines across the full spectrum of values of T . The “Source” baseline exhibits a comparable performance to our method for low values of T , however, its performance barely improves as T increases, in contrast, the “Intrinsic” baseline performs poorly for low values of T but exhibits a comparable performance to our method for high values of T . Finally, “Logistic Regression” achieves an AUC lower than our method across the full spectrum of values of T .

The above results suggest that refutations that occur quickly after a statement is posted are mainly due to the untrustworthiness of the source, while refutations that occur later in time are due to the intrinsic unreliability of the article. As a consequence, our model, by accounting for both source trustworthiness and information intrinsic reliability, can predict both quick and slow refutations more accurately than models based only on one of these two factors.

Can we predict which of the answers to a question in Stack Overflow will be accepted? Unlike Wikipedia where each article receives multiple evaluations (*i.e.*, deleted links), we have only one evaluation (*i.e.*, accepted answer) for every question in Stack Overflow. This property prevents us from estimating question difficulty in the test set and subsequently making predictions similar to that of *Wikipedia*. However, we can estimate users’ reliability from all the questions in the training set and predict which of several competing answers to a question will be most likely verified.

— *Experimental setup:* We first split all questions (and corresponding answers) into a training set (90%

| Stack Overflow | | | |
|----------------|---------|-----------|-----------------------|
| Rank | user-id | ranking | P accept in 4 days |
| 1 | 318425 | top 0.30% | 0.93 |
| 2 | 405015 | top 0.07% | 0.81 |
| 3 | 224671 | top 0.01% | 0.81 |
| 138 | 246342 | top 0.12% | 0.53 |
| 139 | 616700 | top 0.36% | 0.53 |
| Last | 344491 | top 0.97% | 0.53 |

Table 2: *Stack Overflow* users with more than 100 answers (140 users) ranked by the probability that answer they provide is verified within 4 days (Most reliable on top). The table also shows the ranking provided by *Stack Overflow*.

of the questions) and test set (the remaining 10%) at random, then fit the parameters of the evaluation process given by Eq. 9 using only the evaluation times of the answers in the training set, and finally predict which answers will be accepted in the test set by computing the expected verification time for all answers to a question using the fitted model and selecting the earliest estimated verification time. We compare the performance of our model with two baselines: “Source” and “Correct Answers”. “Source” only account for the trustworthiness of the sources (users) and ignores the intrinsic reliability (difficulty) of the questions. Thus, it computes the expected verification time of an answer in the test set as the average verification time of all the answers provided by its associated source user in the training set. Then, for each question in the test set, this baseline selects the answer with the lowest expected verification time. “Correct Answers” ranks sources (users) according to the number of accepted answers posted by each user in the training set. Then, for each question in the test set, it selects the answer with the highest ranked associated source.

— *Results*: Figure 2(b) summarizes the results by means of success rate for different training set sizes. Note that, unlike in the *Wikipedia* experiment, this prediction task does not correspond to a binary classification problem and therefore AUC is not a suitable metric in this case. Our model always achieves a rate of success over 0.47, consistently beats both baselines and, as expected, it becomes more accurate as we feed more events into the estimation procedure. Note that, for most questions, there are more than two answers and the success rate of a random baseline is 0.41. The above results suggest that one needs to account for both the users’ trustworthiness and the difficulty of the questions to be able to accurately predict which answer will be accepted, in agreement with previous work [3].

Do our model parameters provide a meaningful and interpretable measure of source trustworthiness? We answer this question by analyzing the source parameters γ_s and α_s estimated by our parameter estimation method, both in *Wikipedia* and *Stack Overflow*.

First, we pay attention to the 20 most used web sources in *Wikipedia* for two topics, *i.e.*, politics and music, and active users in *Stack Overflow* with over 100 answers, and rank them in terms of source trustworthiness (*i.e.*, in *Wikipedia*, higher trustworthiness means lower α_s while, in *Stack Overflow*, higher trustworthiness means higher α_s). Then, we compute the probability that a statement supported by each source is refuted in less than 6 months in *Wikipedia* or verified in less than 4 days in *Stack Overflow* due to only the source trustworthiness (*i.e.*, setting $\beta = 0$). Table 1 and 2 summarize the results, which reveal several interesting patterns. For example, our model identifies social networking sites such as Twitter, which often accumulate questionable facts and opinionated information, as untrustworthy sources for music in *Wikipedia*. Similarly, for articles related to politics, some notable news agencies close to the left of the political spectrum are considered to be more trustworthy, in agreement with previous studies on political bias in Wikipedia [16]. Moreover, users with high reputation, as computed by *Stack Overflow* itself, are indeed identified in our framework as trustworthy. However, the ranking among these users in terms of reputation do not always match our measure of trustworthiness since it also takes into account other factors such as number of up-votes on questions and answers.

Next, we look at the source parameters at an aggregate level by means of their empirical distribution

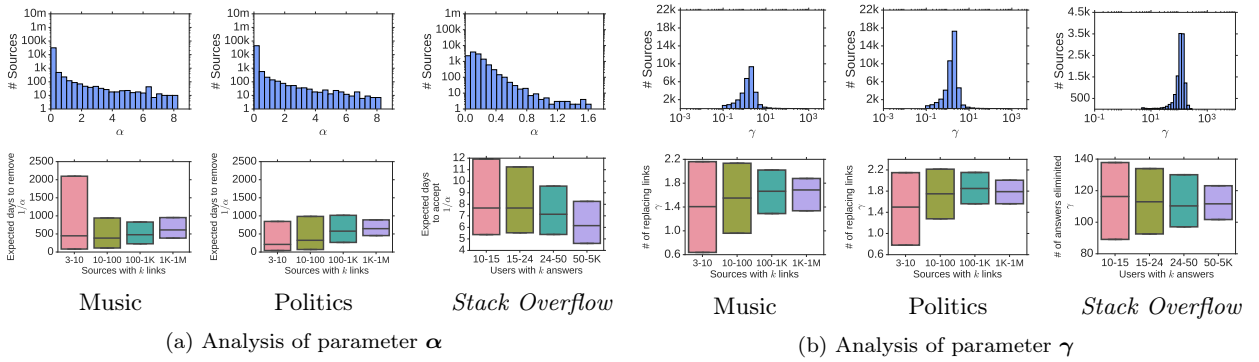


Figure 3: Source Trustworthiness. Panels (a) and (b) show the distributions of the parameters α and γ for the Web sources in *Wikipedia* for the topics “music” and “politics” and for the *Stack Overflow* users, respectively. In both panels, the top row shows the distributions across all sources, while the bottom row shows the distributions for four set of sources, grouped by their popularity in the case of *Wikipedia* and by the number of answered questions in the case of *Stack Overflow* users. In *Wikipedia*, the evaluation mechanism is refutation and thus larger values of $1/\alpha$ correspond to more trustworthy users whose contributed content is refuted more rarely. In *Stack Overflow*, the evaluation mechanism is verification and thus smaller values of $1/\alpha$ correspond to more trustworthy users whose contributed content is verified quicker. In both cases, higher values of γ imply a larger impact on the overall reliability of the knowledge item (*i.e.*, article and question) after an evaluation.

across users. Figure 3 summarizes the results, which show that: (i) the distributions are remarkably alike across both topics in *Wikipedia* and (ii) γ values are distributed similarly both for *Stack Overflow* and *Wikipedia*, however, α values are distributed differently since they capture a different mechanism, verification instead of refutation. Finally, we group web sources in *Wikipedia* by popularity and users of *Stack Overflow* by number of contributed answers, and analyze the source parameters. We summarize the results in Figure 3, which show that: (i) more popular web sources in *Wikipedia* and more active users in *Stack Overflow* tend to be more trustworthy, *i.e.*, lower (higher) α in *Wikipedia* (*Stack Overflow*); (ii) popular sources in *Wikipedia* have a larger impact on the reliability of the article, triggering a larger number of new statements additions (*i.e.*, larger values of γ) after a refutation; and, (iii) there is ample variation across sources in terms of trustworthiness within all groups.

What do the temporal evolution of the intrinsic reliability of Wikipedia articles tell us? In this section, we show that changes on the intrinsic reliability of a *Wikipedia* article closely match external noteworthy events, often controversial, related to the article.

Figure 4 shows the intrinsic reliability both in the statement addition process (first term in Eq. 8), which captures the arrival of new information, and the verification process (first term in Eq. 9), which captures the controversy of the article, for four different articles – Barack Obama’s biography,⁶ George W. Bush’s biography,⁷ an article on 2011 military intervention in Libya,⁸ and an article on the TV show Prison Break.⁹ Each of the articles exhibits different characteristic temporal patterns. In the two biographical articles and the article on the TV show, we find several peaks in the arrival of new information and controversy over time, which typically match remarkable real-world events. For example, in Barack Obama’s article, the peaks in early 2007 and mid-2008 coincide with the time in which he won the Democratic nomination and the 2008 US election campaign; and, in the Prison Break’s article, the peaks coincide with the broadcasting of the four seasons. In contrast, in the article about 2011 military intervention in Libya, we only find one peak, localized

⁶https://en.wikipedia.org/wiki/Barack_Obama

⁷https://en.wikipedia.org/wiki/George_W._Bush

⁸https://en.wikipedia.org/wiki/2011_military_intervention_in_Libya

⁹https://de.wikipedia.org/wiki/Prison_Break

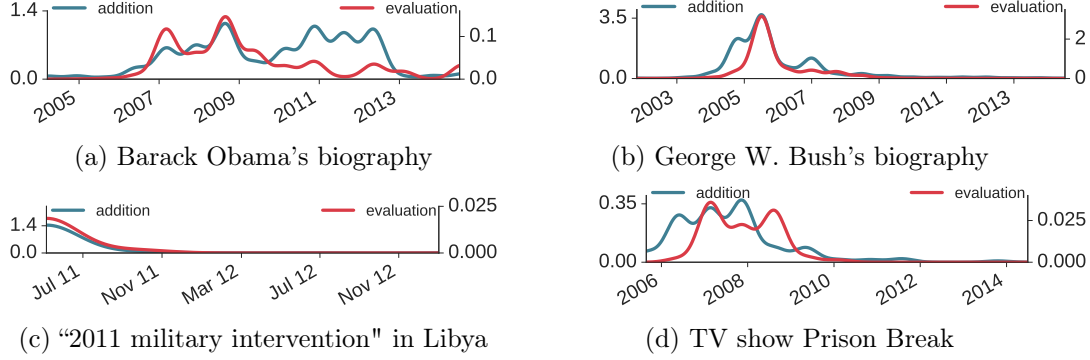


Figure 4: Temporal evolution of the article intrinsic reliability for four *Wikipedia* articles. The blue (red) line shows intensity of statement addition (evaluation) process. Changes on the intrinsic reliability closely match external noteworthy events, often controversial, related to the corresponding article.

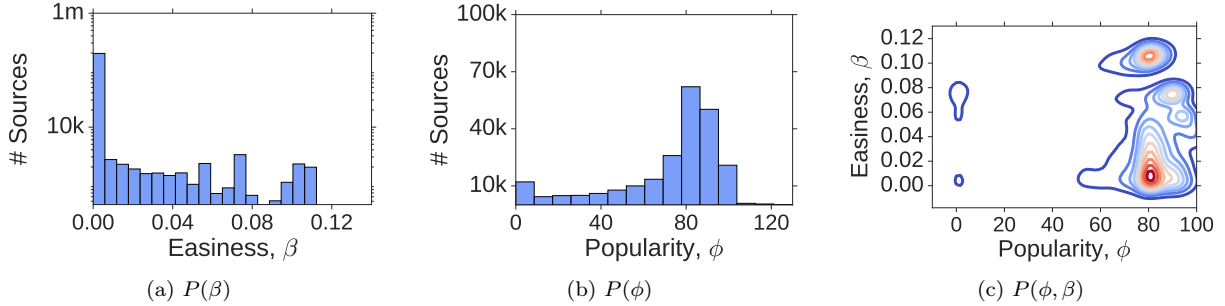


Figure 5: Difficulty vs. popularity in *Stack Overflow* questions. Panels (a) and (b) show the distribution of the parameters β and ϕ , which represent respectively the difficulty and the popularity of *Stack Overflow* questions. Panel (c) shows the joint distribution of both parameters β and ϕ . Higher value of β (ϕ) implies easier (more popular) questions.

at the beginning of the article life cycle, which is followed by a steady decline in which the controversy lasts for a few months longer than the arrival of new information. A comparison of the temporal patterns of new information arrivals and controversy within an article reveals a more subtle phenomenon: while sometimes a peak in the arrival of new information also results in a peak of controversy, there are peaks in the arrival that do not trigger controversy and vice-versa.

What do the intrinsic reliability of *Stack Overflow* questions tell us? We answer this question by analyzing the parameters β_d and ϕ_d estimated by our parameter estimation method for questions in *Stack Overflow*. For each question, such parameters are unidimensional since, unlike *Wikipedia*, the reliability of questions in *Stack Overflow* does not typically change over time. Moreover, the parameters have natural interpretation: β reflects the easiness of a question and ϕ reflects its popularity.

Figures 5(a-b) show the empirical marginal distribution of the parameters across questions and Figure 5(c) shows the joint distribution for questions with $\beta > 0$. The results reveal four clusters: questions which are popular and easy, questions which are popular but difficult, questions that are not popular and difficult, and questions that are not popular but easy.

7 Conclusion

In this paper, we proposed a temporal point process modeling framework of refutation and verification in online knowledge repositories and developed an efficient convex optimization procedure to fit the parameters of our framework from historical traces of the refutations and verifications provided by the users of a knowledge repository. Then, we experimented with real-world data gathered from *Wikipedia* and *Stack Overflow* and showed that our framework accurately predicts refutation and verification events, provides an interpretable measure of information reliability and source trustworthiness, and yields interesting insights about real-world events.

Our work also opens many interesting directions for future work. For example, natural follow-ups to potentially improve the expressiveness of our modeling framework include:

1. Consider sources can change their trustworthiness over time due to, *e.g.*, increasing their expertise [25].
2. Allow for non-binary refutation and verification events, *e.g.*, partial refutations, ratings.
3. Augment our model to consider the trustworthiness of the user who refutes or verifies a statement.

Moreover, we experimented with data gathered from *Wikipedia* and *Stack Overflow*, however, it would be interesting to apply our model (or augmented versions of our model) to other knowledge repositories (*e.g.*, *Quora*), other types of online collaborative platforms (*e.g.*, Github), and the Web at large. Finally, one can think of using our measure of trustworthiness, as inferred by our estimation method, to perform credit assignment in online collaborative platforms—in *Wikipedia*, one could use our model to identify trustworthy users (or dedicated editors) who can potentially make an article more reliable and stable.

References

- [1] O. Aalen, O. Borgan, and H. K. Gjessing. *Survival and event history analysis: a process point of view*. Springer, 2008.
- [2] B. T. Adler and L. De Alfaro. A content-driven reputation system for the wikipedia. In *WWW*, 2007.
- [3] A. Anderson, J. Kleinberg, and S. Mullainathan. Assessing Human Error Against a Benchmark of Perfection. In *KDD*, 2016.
- [4] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5(1):231–297, 2005.
- [5] H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schölkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *ICML*, 2014.
- [6] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. Gomez-Rodriguez. Learning and forecasting opinion dynamics in social networks. In *NIPS*, 2016.
- [7] S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 2016.
- [8] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. From data fusion to knowledge fusion. *VLDB*, 2014.
- [9] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *VLDB*, 2015.
- [10] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. Recurrent Marked Temporal Point Process: Embedding Event History to Vector. In *KDD*, 2016.
- [11] N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha. Scalable influence estimation in continuous-time diffusion networks. In *NIPS*, 2013.

- [12] M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, and L. Song. Shaping social activity by incentivizing users. In *NIPS*, 2014.
- [13] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *NIPS*, 2015.
- [14] M. Farajtabar, X. Ye, S. Harati, L. Song, and H. Zha. Multistage campaigning in social networks. In *NIPS*, 2016.
- [15] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *ICML*, 2011.
- [16] S. Greenstein and F. Zhu. Is wikipedia biased? *The American economic review*, 102(3):343–348, 2012.
- [17] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB*, 2004.
- [18] D. Hunter, P. Smyth, D. Q. Vu, and A. U. Asuncion. Dynamic egocentric models for citation networks. In *ICML*, 2011.
- [19] M. Karimi, E. Tavakoli, M. Farajtabar, L. Song, and M. Gomez-Rodriguez. Smart Broadcasting: Do you want to be seen? In *KDD*, 2016.
- [20] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. On the discovery of evolving truth. In *KDD*, 2015.
- [21] X. Liu, X. L. Dong, B. C. Ooi, and D. Srivastava. Online data fusion. *VLDB*, 2011.
- [22] A. Pal, V. Rastogi, A. Machanavajjhala, and P. Bohannon. Information integration over time in unreliable and uncertain environments. In *WWW*, 2012.
- [23] J. Pasternack and D. Roth. Latent credibility analysis. In *WWW*, 2013.
- [24] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *LREC*, 2010.
- [25] U. Upadhyay, I. Valera, and M. Gomez-Rodriguez. Uncovering the dynamics of crowdlearning and the value of knowledge. In *WSDM*, 2017.
- [26] I. Valera and M. Gomez-Rodriguez. Modeling adoption and usage of competing products. In *ICDM*, 2015.
- [27] S. Wang, D. Wang, L. Su, L. Kaplan, and T. F. Abdelzaher. Towards cyber-physical systems in social spaces: The data reliability challenge. In *RTSS*, 2014.
- [28] M. Wu and A. Marian. Corroborating answers from multiple web sources. In *WebDB*, 2007.
- [29] H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng., and A. Zhang. Towards confidence in the truth: A bootstrapping based truth discovery approach. In *KDD*, 2016.
- [30] X. Yin and W. Tan. Semi-supervised truth discovery. In *WWW*, 2011.
- [31] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. *Proceedings of QDB*, 2012.
- [32] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *VLDB*, 2012.
- [33] K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional hawkes processes. In *ICML*, 2013.